DOCUMENT RESUME

ED 174 680 **TM 009 652**

AUTHOR Smith, Douglas G.
TITLE Estimating the Reliability of Clinical Ratings: An Alternative Approach.
PUB DATE Apr 79
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Computer Programs; Higher Education; Informal Assessment; *Mathematical Formulas; *Medical Students; *Rating Scales; *Reliability; Statistical Analysis; *Student Evaluation
IDENTIFIERS Interrater Reliability; Kappa Coefficient; *Rater Reliability; *Reliability Formulas

ABSTRACT

The ratings of 43 third-year medical students were used to investigate a method of computing the reliability of ratings. The students were evaluated on a four-point rating scale along seven dimensions of clinical ability: diagnostic ability; motivation; sense of responsibility; oral-verbal ability; effectiveness with patients; ability to take a history and perform a physical examination; and cooperation. Seven individuals rated these students. The ratings were analyzed by means of a computer program which used a pairwise definition of agreement, and provided the following information: (1) distribution of ratings for each rater; (2) number of agreements as well as the proportion of times the rater was in agreement with other raters with whom he was paired; (3) kappa coefficient (corrected for chance agreement) for each rater on each dimension; (4) maximum value of kappa given the marginal distributions of ratings; and (5) mean and standard deviation of ratings. This information was also summarized for each rater across all dimensions, as well as for all raters on each dimension. An analysis of variance estimate of reliability was computed for comparison. The author's method provided greater meanings for the estimates, as well as an indication of the sources of unreliability. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SESSION 3.09

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED174680

Estimating the Reliability of Clinical Ratings: An Alternative Approach

Douglas U. Smith

Division of Educational Services and Research
Wayne State University School of Medicine
Detroit, Michigan 48202

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Douglas U. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California, April 1979.

TM009 652

Estimating the Reliability of Clinical Ratings: An Alternative Approach

Introduction

Rating scales are frequently employed throughout a physician's training in order to assess performances which would be difficult, if not impossible, to measure, say, by means of multiple-choice examinations. However, as with any type of assessment instrument, care must be taken to carefully monitor the quality of measurement. This paper describes a method for assessing the reliability of ratings which is conceptually simpler, easier to compute, more directly interpretable, and which provides more information regarding rater performance than the more commonly used analysis of variance (ANOVA) model (Hoyt, 1941; Ebel, 1951).

Inherent in the use of rating scales to assess the clinical performances of medical students are a number of problems which, in other types of educational settings, can be adequately controlled by means of careful design considerations. These problems include the likelihood that: (1) different students will be rated by different raters, (2) students will receive different numbers of ratings, and (3) raters will rate different numbers of students. These problems along with the natural variations which exist among raters (see Guilford, 1954), makes the task of insuring reliable ratings much more difficult.

It is the purpose of this paper to describe an alternative approach to the estimation of the reliability of clinical ratings which does not require either "ideal circumstances" or extraordinary assumptions regarding the nature of between-rater differences. Additionally, this method provides much more information with respect to rater performance than does the more conventional ANOVA approach. In short, the proposed method involves computing a simple agreement statistic (corrected for chance) between pairs of raters.

Several advantages accrue from the implementation of such a method. These include: (1) the measurement of rater agreement (reliability) for each rater on each trait to be assessed, (2) a similar measure of agreement for each rater across all traits, (3) a measure of rater agreement for all raters on a single trait, and finally, (4) an overall measure of rater agreement for all raters on all traits. In terms of practicality, it becomes possible to use this type of information to identify the particular strengths and weaknesses of individual raters, as well as to aid in the development and evaluation of programs designed to improve the observational skills of the raters.

Background

Any review of the literature regarding the reliability of ratings would, of necessity, require more space than is available here. There exist, however, a number of articles of particular interest. Included among these are: Hoyt (1941), Ebel (1951), and Stanley (1963).

Traditionally, the estimation of the reliability of ratings has been

carried out by means of analysis of variance (ANOVA) methods. Ebel (1951) discussed two measures of reliability which could be derived from these procedures. One basically answers the question, "How reliable are the individual ratings?" and the other, the question, "How reliable are the average ratings?" Choice of either one of these coefficients depends on a number of factors. First, if several persons are responsible for rating a student, and if an individual's total, or average rating is used as the basis for assessing performance, then it is necessary to consider estimating the reliability of these average ratings. However, when the raters operate independently of one another or when "multiple scores are only available in experimental situations", then reliability estimates based on individual ratings are most appropriate (Ebel, 1951, p. 117).

These methods, however, do not effectively deal with what Guilford (1954) has termed "constant errors" (e.g. errors of leniency, central tendency, and halo effects). When corrections for these types of errors are necessary, one should consult Guilford's text. However, this approach may not be completely satisfactory when one is dealing with incomplete matrices of ratings.

An alternative method of assessing the reliability of ratings is through the use of some type of measure of rater agreement. As a crude index, one might consider using the proportion of times two raters were in agreement with each other. A number of investigators (Hoffman & Overall, 1976; Cohen, 1960; Fleiss et al., 1972, and Light, 1971) have indicated that this approach fails to take into consideration those agreements which could have occurred on a chance basis. Cohen (1960) proposed a coefficient of agreement, kappa, which

in effect, takes into account chance agreement. Kappa will always range from -1 (perfect disagreement) to +1 (perfect agreement) when the marginal distributions of ratings are identical. A value of zero indicates agreement at the chance level. When the marginal distributions for any pair of raters are not the same, the maximum value of kappa will be less than one. In such cases, it is always best to consider the ratio of kappa to the maximum value kappa can assume given the marginal distributions as has been suggested by Cohen (1960).

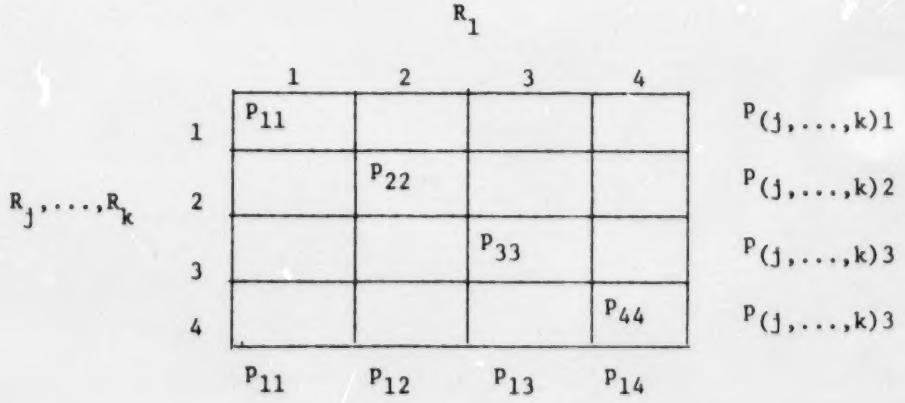
Other methods of assessing rater agreement have also been proposed. One, a modification of kappa, referred to as weighted kappa (Cohen, 1968), allows for situations in which the degree of disagreement is taken into consideration. The unweighted kappa coefficient, in fact, has been shown to be a special case of weighted kappa when all disagreements are assigned the same weight. Several problems exist with the use of weighted kappa. The first involves the choice of weights to be applied to the possible categories of disagreement, while the second involves the interpretation of this coefficient (Hoffman and Overall, 1976).

Still other measures of rater agreement may be used. Light (1971) noted that when only two categories of assignment are possible, then chi-square type measures such as the phi coefficient may be used. However, if the number of categories of assignment exceeds two, then "standard measures of association may be inappropriate for measuring agreement" (Light, 1971, p. 366).

Recent work by Hubert (1977) extends the applicability of the unweighted

kappa coefficient to situations involving the multiple raters and using any one of several concepts of agreement. Germane to the present paper is a "pairwise" definition of agreement which can be formally stated; agreement occurs if and only if two raters classify an individual similarly.

Extending this notion to situations in which there are multiple raters, it becomes possible to obtain for each rater (or, perhaps a group of raters) a kappa-like coefficient describing the extent to which he is in agreement with other raters with whom he was paired. If we consider for each rater pair a T X T matrix, where T represents the possible ratings which a student may receive, and for a particular rater R_1, there exist a total of k (one for each time R_1 is paired with another rater) such matrices, then we can collapse these separate matrices into a single T X T matrix which summarizes the agreement between R_1 and all other raters with whom he was paired. This is illustrated in the following diagram.



In this example, T has four levels and $R_j,\ldots,R_k$ refers to the other raters who have been paired with R_1. The entries along the main diagonal represent the observed proportion of agreements for each level of T. In order to compute

the kappa coefficient for R_1, the expected proportion of agreements must be determined. Expected proportions of agreement are found for each level of T by multiplying the corresponding row and column marginal proportions. Their sum represents the expected proportion of agreements. Once this has been determined, kappa can be found by substitution of the observed proportion of agreements, p_o, and the expected proportion of agreements p_c, into the following equation: $\kappa = (p_o - p_c)/(1 - p_c)$.

Further application of this approach also allows for a description of rater agreement for all traits as well as for all raters on a particular trait of interest. The following section describes an application of this approach in to assessing the reliability of clinical ratings.

Method

The clinical ratings of 43 third-year medical students in a Gynecology/Obstetrics rotation were used in the present analysis. This group of students had approximately 24 weeks of prior experience in other clinical clerkships.

The students were evaluated by means of a four-point rating scale along seven dimensions of clinical ability. These dimensions were: (1) diagnostic ability (DIAG), (2) motivation (MOTI), (3) sense of responsibility (RESP), (4) oral-verbal ability (OVAB), (5) effectiveness with patients (EFFT), (6) ability to take a history and perform a physical examination (H&PH), and (7) cooperation with personnel (COOP). Seven individuals were assigned to rate these students.

Student ratings were then analyzed by means of a computer program written

by this investigator (Smith, 1978). Using a pairwise definition of agreement, this program provides the following information: (1) the distribution of ratings for each rater, (2) the number of agreements as well as the proportion of times the rater was in agreement with other raters with whom he was paired (PAGR), (3) a kappa coefficient (corrected for chance agreement) for each rater on each dimension (KAPPA), (4) the maximum value of kappa given the marginal distributions of ratings (KMAX), and (5) the mean and standard deviation (STDEV) of ratings. Additionally, this information is summarized for each rater across all dimensions as well as for all raters on each dimension.

Results

Table 1 displays a summary of the information which is provided for each rater, while Table 2 describes the performance of all raters within the GYN/OB clerkship. In addition to the information discussed above, these tables also present the number of times the rater(s) did not observe a particular trait or ability (N.O.), the number of agreements (NAGR) and the number of rater pairings (NPAIR). Also, it should be noted that rater performance has been summarized by collapsing across all of the traits for a single rater (TOTL) as well as across all raters (Table 2, TOTL).

Insert Table 1 about here

In Table 2 an analysis of variance (Hoyt, 1941) estimate of the reliability of ratings for each trait has been included for comparative purposes. It should be mentioned that this coefficient (REL) is sometimes interpreted as the proportion of scores which agree, however, such an interpretation is not correct.

Insert Table 2 about here

Discussion

Since classical methods of estimating the reliability of ratings are frequently uninformative to the medical educator, other than in a general sense, an alternative approach based on a pairwise definition of rater agreement was employed. This approach not only seems to provide greater meaning for the estimation of the reliability of ratings, but also a means by which sources of unreliability may be detected. In particular the present methodology allows for the identification of raters for whom the definition of the scales (i.e. "behavioral boundaries") might not be clear.

For whatever reasons these problems occur, a mechanism does exist in order to identify the possible sources of unreliability in clinical rating scales. Identification of these sources of error can be extremely useful in improving the overall quality of the clinical evaluation of medical students.

References

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Fleiss, J. L. Measuring nominal scale agreement among raters. Psychological Bulletin, 1971, 76, 378-382.

Guilford, J. P. Psychometric methods. New York: McGraw -Hill, 1954.

Hoffman, N.G., & Overall, J.E. Problems in the interpretation of indices of diagnostic agreement. Psychometric Lab Reports, No. 59, Galveston, Texas: The University of Texas Medical Branch, Dec. 1976.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Hubert, L. Kappa revisted. Psychological Bulletin, 1977, 84, 289-297.

Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychological Bulletin, 1971, 76, 365-377.

Smith, Douglas U. Assessing rater performance in the evaluation of clinical competencies of third-year medical students and residents: A FORTRAN IV computer program. Detroit, Mich.: Wayne State University School of Medicine.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), Educational measurement (2nd Edition). Washington, D.C.: American Council on Education, 1968.

Table 1

Sample Rater Performance Summary

Rater ID	Trait	Scale Values 1	2	3	4	N.O.	PAGR	KAPPA	KMAX	MEAN	STDEV
846	DIAG	0 (0.0)	1 (0.03)	11 (0.34)	20 (0.63)	0 (0.0)	.70 NAGR=	.31 (49)	.90 NPAIR=	3.59 (70)	.56
	MOTI	0 (0.0)	3 (0.09)	6 (0.19)	23 (0.72)	0 (0.0)	.68 NAGR=	.29 (52)	.65 NPAIR=	3.63 (76)	.66
	RESP	0 (0.0)	1 (0.03)	23 (0.72)	8 (0.25)	0 (0.0)	.49 NAGR=	.15 (37)	.34 NPAIR=	3.22 (76)	.49
	OVAB	0 (0.0)	0 (0.0)	19 (0.59)	13 (0.41)	0 (0.0)	.72 NAGR=	.45 (53)	.66 NPAIR=	3.41 (74)	.50
	EFFT	0 (0.0)	1 (0.03)	10 (0.31)	21 (0.66)	0 (0.0)	.79 NAGR=	.52 (33)	.79 NPAIR=	3.63 (42)	.55
	H&PH	0 (0.0)	0 (0.0)	10 (0.31)	22 (0.69)	0 (0.0)	.77 NAGR=	.11 (30)	.31 NPAIR=	3.69 (39)	.47
	COOP	0 (0.0)	0 (0.0)	9 (0.28)	23 (0.72)	0 (0.0)	.82 NAGR=	.39 (36)	.85 NPAIR=	3.72 (44)	.46
RATER	TOTL	0 (0.0)	6 (0.03)	88 (0.39)	130 (0.58)	0 (0.0)	.69 NAGR=	.35 (290)	.80 NPAIR=	3.55 (421)	.55

Table 2

Sample Performance Summary -- All Raters

Trait	1	2	3	4	N.O.	REL	PAGR	KAPPA	XMAX	MEAN	STDEV
DIAG	0	1	37	74	30	.58	.69	.24	.94	3.65	.50
	(0.0)	(0.01)	(0.26)	(0.52)	(0.21)		NAGR=	(82)	NPAIR=	(119)	
MOTI	0	7	33	78	24	.66	.65	.2[illegible]	.82	3.60	.60
	(0.0)	(0.05)	(0.23)	(0.55)	(0.17)		NAGR=	(90)	NPAIR=	(138)	
RESP	0	4	55	59	24	.34	.53	.10	.93	3.47	.57
	(0.0)	(0.03)	(0.39)	(0.42)	(0.17)		NAGR=	(73)	NPAIR=	(138)	
OVAB	0	0	56	60	26	.57	.64	.29	.96	3.52	.50
	(0.0)	(0.0)	(0.39)	(0.42)	(0.18)		NAGR=	(85)	NPAIR=	(132)	
EFFT	0	1	31	52	58	.64	.77	.46	.75	3.61	.52
	(0.0)	(0.01)	(0.22)	(0.37)	(0.41)		NAGR=	(44)	NPAIR=	(57)	
H&PH	0	0	17	64	61	.50	.80	.14	.14	3.79	.41
	(0.0)	(0.0)	(0.12)	(0.45)	(0.43)		NAGR=	(40)	NPAIR=	(50)	
COOP	0	0	21	64	57	.61	.79	.29	.40	3.75	.43
	(0.0)	(0.0)	(0.15)	(0.45)	(0.40)		NAGR=	(49)	NPAIR=	(62)	
TOTL	0	13	250	451	280		.67	.27	.88	3.61	.52
	(0.0)	(0.01)	(0.25)	(0.45)	(0.28)		NAGR=	(463)	NPAIR=	(696)	